Data Compression and Digital Trees

W. Szpankowski
Purdue University

October 5, 1992

[summary by Mireille Régnier]

Abstract

In this talk, the author shows the relationship between two classical data compression algorithms due to Lempel and Ziv [9], and well-known data structures, namely Digital Search Trees and Suffix Trees. Hence, the performance evaluation of these data compression algorithms reduces to the analysis of some tree parameters. Second-order properties are derived. A normal limiting distribution is conjectured. Also, some open problems are given.

1. Introduction

Section 2 briefly presents data compression and the relationship to trees. Then, in Section 3, we study the relationship between tree parameters and data compression performance; we formulate some mathematical problems. Section 4 deals with second order properties, and notably describes the approach of the author. Finally, we provide in Section 5 a small list of open problems. Most of these results appear in INRIA research report [11] and in [6].

2. Lempel and Ziv data compression algorithms

The data compression problem is the following. Some "known" string of length n, the so-called database, is given. One must find the longest substring of the database string that is identical to a yet "unknown" sequence (to be manipulated). Lempel and Ziv algorithms realize a partition of the database sequence into blocks. This is called *parsing*. The parsing satisfies the following properties:

(i) blocks are pairwise distinct;
(ii) each block that occurs in the parsing has already been seen somewhere to the left.

Example. Let us consider sequence

$$11001010001000100$$

In the first algorithm, LZ1, overlapping is not allowed, that is a previous occurrence is not taken into account if it is shared between two consecutive occurrences. Overlapping is allowed in the second one, LZ2. This leads to the two partitions:

$$(1)(10)(0)(101)(00)(01)(000)(100) \ : \ LZ1$$

and

$$(1)(10)(0)(101)(00)(01)(000100) \ : \ LZ2$$

Note that difference occur at position 12. Sequence 000 occurred before, but is split between blocks 5 and 6.

LZ1 is associated to a digital search tree built on the block sequence read from left to right. LZ2 is associated to a suffix tree. We present on Figure 2 the tree associated to LZ1.

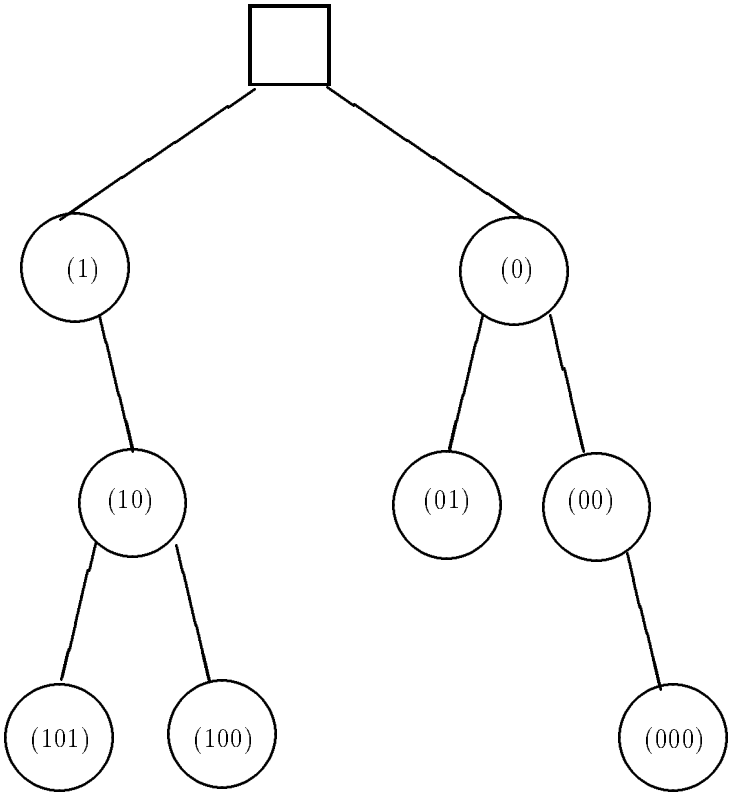

FIGURE 1. A digital tree representation of Ziv's parsing for the string 11001010001000100 ...

3. Relevant parameters

The relevant parameters for data compression are:

- M_n: number of phrases built over a (random) sequence of length n;
- $M_n(k)$: number of phrases of length k;
- l_m: length of the m-th phrase;
- H_n: length of the longest phrase;
- $\tilde{I}_n$: longest substring that can be duplicated;
- N_l: size of a database having two copies of some substring of length l.

The relevant parameters for a digital search tree are given below. We consider a tree built over n strings.

- $D_n(m)$: depth of the m-th string;
- D_n: depth of a randomly selected string:

$$Pr(D_n \leq k) = \frac{1}{n} \sum_{m=1}^{n} Pr(D_n(m) \leq k);$$

- $I_n = D_{n+1}(n+1)$: depth of insertion;
- H_n: height of the tree, i.e. $\max_{1 \leq m \leq n}(D_n(m))$;
- S_n: shortest path in the tree, e.g. $\min_{1 \leq m \leq n}(D_n(m))$;
- L_n : external pathlength

$$L_n = \sum_{m=1}^{n} D_n(m);$$

- Z_n : size of the tree, i.e. number of internal nodes.

We now provide a list of relationships between data compression and tree parameters.

Non-overlapping parsing algorithm LZ1.

$$Z_{M_n} = M_{n+1}, \tag{1}$$

$$L_{M_n-1} < n \leq L_n, \tag{2}$$

$$l_m = D_m(m) = I_m, \tag{3}$$

$$M_n(k) = \#\text{ of internal nodes at level}k. \tag{4}$$

Parsing algorithm LZ2.

$$l_k = I_{\sum_{r=1}^{k-1} l_r} = D_{\sum_{r=1}^{k-1} l_r}(\sum_{r=1}^{k-1} l_r), \tag{5}$$

$$\tilde{I}_n = I_n = D_n(n), \tag{6}$$

$$D_{N_l}(1) = l. \tag{7}$$

Unfortunately, there exists no simple relationship between M_n and L_n. We only have:

$$\sum_{k=1}^{M_n} l_k = n.$$

4. Deriving limiting distributions

First order properties, i.e. average values, convergence in probability or almost sure have now been derived for many classes of trees. See notably [1, 8, 10] for Digital Search Trees and [13] for suffix trees.

Second order properties, e.g. variances, large deviation results and limiting distributions are less known. Most results are derived for tries [2, 4, 5, 7]. Digital Search Trees were open.

The first result presented is the limiting distribution of the number of phrases of length k, in LZ1, i.e. $M_n(k)$ or D_n. One proves:

THEOREM 1. *(i) For the symmetric Bernoulli model the limiting distribution of D_m is*

$$\lim_{m\to\infty} \Pr\{D_m = x + \log_2 m\} = 2^{x-1}\left(1 + \frac{1}{Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{-2^{-i(i+1)/2}}{Q_i}e^{-2^{-(x-1-i)}}\right) \tag{8}$$

for such real x that $x+\log_2 m$ is integer, with $Q_k = \prod_{j=1}^{k}(1-2^{-j})$. (ii) In the asymmetric case, the limiting distribution of D_m is normal, that is,

$$\frac{D_m - ED_m}{\sqrt{\text{Var } D_m}} \to N(0,1) \tag{9}$$

where ED_m and Var D_m *are given by (10) and (11), respectively.*

$$ED_m = \frac{1}{h}\left(\log m + \gamma - 1 + \frac{H}{2h} + \theta + \delta(m)\right) + O(\log m/m) \tag{10}$$

$$\text{Var } D_m = \frac{H - h^2}{h^3}\log m + A + \Delta(m) + O(\log^2 m/m) \tag{11}$$

Moreover, the moments of D_m converges to the appropriate moments of the normal distribution. More precisely, for any complex ϑ

$$e^{-\vartheta c_1 \log m}E(e^{\vartheta D_m}) = e^{c_2\frac{\vartheta^2}{2}\log m}\left(1 + O(\frac{\vartheta}{\sqrt{\log m}})\right) \tag{12}$$

where $c_1 = 1/h$ *and* $c_2 = (H - h^2)/h^3$.

PROOF. One considers the generating functions $B_n(z)$, where $[z^k]B_n(z)$ is the average number of internal nodes at level k. One proves the recurrence equation:

$$B_m(u) = m - (1-u)\sum_{k=2}^{m}(-1)^k \binom{m}{k} Q_{k-2}(u) \tag{13}$$

where

$$Q_k(u) = \prod_{j=1}^{k}(1 - u2^{-j}). \tag{14}$$

Since the formula for $Q_k(u)$ is relatively simple, we can extract coefficients of $B_m(u)$ "by hand".

Note that $Q_k(u) = Q_\infty(u)/Q_\infty(u2^{-k})$, and, as in Louchard [10],

$$\frac{1}{Q_\infty(u)} = \sum_{i=0}^{\infty}\frac{u^i}{2^i Q_i}, \quad Q_\infty(u) = -\sum_{i=0}^{\infty} u^i R_i \tag{15}$$

where

$$R_i = (-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i} \tag{16}$$

with $Q_i = Q_i(1)$. Let now $[u^k]f(u)$ denote the coefficient at u^k of $f(u)$. Note that

$$[u^n]Q_{k-2}(u) = -\sum_{l=0}^{n}\frac{R_{n-l}}{Q_l 2^{l(k-1)}}.$$

Hence applying this to our basic solution (13) we obtain

$$\begin{aligned}[u^{j+1}]B_m(u) &= \sum_{l=0}^{j+1}\frac{2^l R_{j+1-l}}{Q_l}\left((1-2^{-l})^m - 1 - m/2\right) \\ &- \sum_{l=0}^{j}\frac{2^l R_{j-l}}{Q_l}\left((1-2^{-l})^m - 1 - m/2\right).\end{aligned}$$

Finally, after some tedious algebra one obtains an explicit formula as in Louchard [10], and taking $m \to \infty$ we easily derive part (i) of Theorem 1 (see also Mahmoud [12], Ex. 6.12).

Alternatively, Mellin-like or Rice method can be used to find an asymptotic solution. Then, one may use Cauchy formula to extract coefficient $M_n(k)$. In the asymmetric case, one considers $D_m(u) = B_m(u)/m$. □

We come now to the second result: the limiting distribution of the number of phrases M_n in LZ1.

THEOREM 2. *(i) The length of a randomly selected phrase for the symmetric Bernoulli model has the following limiting distribution*

$$\lim_{n\to\infty}\Pr\{D_n^{LZ} = x + \log_2(n/\log_2 n)\} = 2^{x-1}\left(1 + \frac{1}{Q_\infty}\sum_{i=0}^{\infty}(-1)^{i+1}\frac{2^{-i(i+1)/2}}{Q_i}e^{-2^{-(x-1-i)}}\right) \tag{17}$$

for such real x *that* $x + \log_2(n/\log_2 n) = j$ *is an integer.*

(ii) For the asymmetric Bernoulli model the typical depth D_n^{LZ} *is normally distributed. More precisely,*

$$\frac{D_n^{LZ} - c_1\log(nh/\log n)}{\sqrt{c_2\log(nh/\log n)}} \to N(0,1) \tag{18}$$

provided our Conjecture is true. In fact, the rate of convergence is $1 + O(1/\sqrt{\log n})$.

PROOF. The author expresses this random variable D_n^{LZ} as a function of the r.v. D_m of the previous theorem. □

The third result is about external path length. In [6], it is proven that:

THEOREM 3. *Consider a digital search tree under the asymmetric Bernoulli model. Then,*

$$\frac{L_m - c_1 m \log m}{\sqrt{c_2 m \log m}} \to N(0,1), \tag{19}$$

where $c_1 = 1/h$ *and* $c_2 = (H - h^2)/h^3$ *with* $h = -p\log p - q\log q$ *being the entropy of the alphabet and* $H = p\log^2 p + q\log^2 q$. *That is, for* x *real we have* $\lim_{m\to\infty} \Pr\{L_m < c_1 m\log m + x\sqrt{c_2 m \log m}\} = 1/\sqrt{2\pi}\int_{-\infty}^{x} e^{-t^2/2}dt$. *Moreover,*

$$EL_m = c_1 m \log m + O(m) \tag{20}$$

$$\text{Var } L_m = c_2 m \log m + O(m), \tag{21}$$

and all moments of L_m *converge to the appropriate moments of the normal distribution. In the symmetric case (i.e.,* $p = q = 0.5$*), the internal path length* L_m^{sym} *still satisfies (19) with* $EL_m^{sym} \sim \log_2 m$ *and*

$$\text{Var } L_m^{sym} \sim (C + \delta(\log_2 m))m \tag{22}$$

where $C = 0.26600\ldots$ *and* $\delta(x)$ *is a fluctuating continuous function with period 1 (cf.* [8]*). In this case, the convergence in moments also holds.*

PROOF. The scheme of the proof is similar to [3]: the bivariate generating function for the external path length is defined by two equations:

$$L_{m+1}(u) = u^m \sum_{k=0}^{m} \binom{m}{k} p^k q^{m-k} L_k(u) L_{m-k}(u). \tag{23}$$

with $L_0(u) = 1$. Hence, also

$$\frac{\partial L(z,u)}{\partial z} = L(pzu, u)L(qzu, u) \tag{24}$$

with $L(z, 0) = 1$.

(1) one first analyzes the Poisson model that is characterized by the exponential bivariate generating function $L(z, u)$ satisfying (24).
(2) In order to solve (24) one tries to transform it into an additive functional equation by considering $\log L(z, u)$. This is only possible if one proves the existence of $\log L(z, u)$. Hence, one proves that there is a convex cone around the real axes such that for some $\kappa(u)$ we have $\log L(z, u) = \Theta(z^{\kappa(u)})$.
(3) Next, one uses Taylor expansion of $\log L(z, u)$ in the convex cone to show that for large z the generating function $L(z, u)$ appropriately normalized converges to the generating function of the normal distribution.
(4) The final effort is to de-Poissonize the latter result, that is, to transform the normal distribution of the Poisson model into the normal distribution of the Bernoulli model.

□

5. Open problems

Many problems remain open. One would like to extend first and second order results in the case of Markovian distributions. Also, the variance of the external path length is of interest for asymmetric Bernoulli model and various classes of trees: Digital Search Trees, tries, Patricia tries.

Bibliography

[1] Flajolet (P.) and Sedgewick (R.). – Digital search trees revisited. *SIAM Journal on Computing*, vol. 15, n° 3, August 1986, pp. 748–767.

[2] Jacquet (Ph.) and Régnier (M.). – Normal limiting distribution of the size of tries. In Courtois (P.-J.) and Latouche (G.) (editors), *Performance'87*. pp. 209–223. – North-Holland, 1987. 12-th IFIP WG International Symposium on Computer Performance, Bruxelles.

[3] Jacquet (Ph.) and Régnier (M.). – *Normal limiting distribution of the size and external path length of tries.* – Research report n° 827, Institut National de Recherche en Informatique et en Automatique, 1988.

[4] Jacquet (Ph.) and Régnier (M.). – New results on the size of tries. *IEEE Transactions on Information Theory*, vol. 35, n° 1, 1989, pp. 203–205.

[5] Jacquet (Ph.) and Szpankowski (W.). – Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, vol. 37, 1991, pp. 669–675.

[6] Jacquet (Ph.) and Szpankowski (W.). – *A Functional Equation Arising in the Analysis of Algorithms.* – Technical report, Institut National de Recherche en Informatique et en Automatique, 1993.

[7] Jacquet (Philippe) and Régnier (Mireille). – Trie partitioning process: Limiting distributions. In Franchi-Zanetacchi (P.) (editor), *CAAP'86*, *Lecture Notes in Computer Science*, volume 214, pp. 196–210. – 1986. Proceedings of the 11th Colloquium on Trees in Algebra and Programming, Nice France, March 1986.

[8] Kirschenhofer (P.), Prodinger (H.), and Szpankowski (W.). – Digital search trees again revisited: The internal path length perspective. *SIAM Journal on Computing*, 1993.

[9] Lempel (A.) and Ziv (J.). – On the complexity of finite sequences. *IEEE Transactions on Information Theory*, vol. 22, n° 1, 1976, pp. 75–81.

[10] Louchard (G.). – Exact and asymptotic distributions in digital and binary search trees. *RAIRO Theoretical Informatics and Applications*, vol. 21, n° 4, 1987, pp. 479–495.

[11] Louchard (G.) and Szpankowski (W.). – *Average Profile and Limiting Distribution for a Phrase Size in the Lempel-Ziv Parsing Algorithm.* – Research Report n° 1886, Institut National de Recherche en Informatique et en Automatique, 1993.

[12] Mahmoud (Hosam). – *Evolution of Random Search Trees.* – New York, John Wiley, 1992.

[13] Szpankowski (W.). – A generalized suffix tree and its (un)expected behaviors. *SIAM Journal on Computing*, 1993.